X Congresso Brasileiro de Inteligência Computacional (CBIC'2011), 8 a 11 de Novembro de 2011, Fortaleza, Ceará

On the use of Lempel-Ziv complexity in signal detection

Jugurta Montalvão, Jânio Canuto
Universidade Federal de Sergipe (UFS), Universidade Estadual de Campinas (UNICAMP)
jmontalvao@ufs.br, jcanuto@decom.fee.unicamp.br

Abstract – In this paper, the seminal method proposed by Abraham Lempel and Jacob Ziv, aimed at the complexity analysis of a single symbolic sequence, was modified to compare similarities between two sequences. This modification allowed the creation of a new criterion which can replace likelihood in some pattern recognition applications. Moreover, to allow for analysis and comparison of multivariate continuously valued patterns, we also present a simple adaptation of the Lempel-Ziv's method to time-sampled signals. To illustrate the usefulness of these proposed tools, experimental results are presented on healthcare signal detection.

Keywords – Signal detection; No *a priori*; Lempel-Ziv.

1 INTRODUCTION

In 1976, A. Lempel and J. Ziv [1] proposed an approach for complexity analysis of symbol sequences. An important aspect of their approach is the lack of *a priori* with regard to the source of symbols, which clearly contrasts with the measurement of (source) Shannon entropy [2]. Though they are conceptually different measures, it was shown that [3], under ergodicity conditions, Lempel-Ziv's complexity of increasingly long symbol sequences converges almost surely to the Shannon entropy of the source from which symbols are drawn.

Lempel-Ziv's (LZ) approach, latterly simplified for practical reasons, became widely known as the compression algorithm behind many computer programs for file compression - the "zip-like" programs. We should probably credit its success to its universality, in other words, to its lack of *a priori*. Nevertheless, it should also be highlighted that zip-like programs are just the "tip of the iceberg," for compression is just a single offspring of the elegant theory presented in [1].

Indeed, compression is a consequence of redundancy removal, and "zip-like" programs aim at finding redundancies in streams of symbols, regardless of what they represent (e.g. texts, audio, video). These redundancies may even be gathered in a so-called "dictionary", whose content corresponds to unique non-overlapping segments of the analyzed sequence of symbols. We highlight though that the idea of an explicit dictionary does not take part in the original analysis method proposed in [1]. Therefore, the dictionary definition in this paper is arbitrary.

The original method by Abraham Lempel and Jacob Ziv was aimed at symbolic sequence analysis, but it may be adapted to work with sequences of numbers as well (e.g. sampled signals). For instance, this adaptation can be done rather straightforwardly through simple quantization of the signal, thus mapping it back into a sequence of labels (one label per quantization level). In all cases, redundancy analysis allows for:

- compression
- segmentation
- pattern recognition
- prediction

In this brief paper, we propose a new method which can be regarded as a tool for pattern recognition. Signal comparison through LZ approach is not new, though not too current in the pattern recognition scenario. One rather isolated example (to the extent of the authors' knowledge), published in 1995, is the work by P. Johansen [4], in which handwritten signature authenticity is verified (behavioural biometrics). Indeed, the asymmetric measure defined in Section 3 is closely related to the ideas presented in [4]. In [5], it was clearly shown, through numeric examples, that LZ complexity may replace the Lyapunov exponent as a more precise measure of order/disorder, in spatiotemporal pattern analysis. Besides, in [6] the Lempel-Ziv-Welch algorithm was successfully applied to texture (image) classification.

In spite of the above mentioned examples, we also observe that LZ based signal analysis is more common in specific research domains, such as biomedical signals [7–10], possibly due to the nice properties of the LZ-based entropy estimators, as compared to the plug-in method [7, 11], along with its simplicity of use. Moreover, because LZ algorithms only process sequences of symbols, thresholding and labeling are frequently applied to real-valued signals prior to LZ analysis. Indeed, in most cases, a single threshold is used to generate two-symbol sequences. For instance, in [4] binary pixel attributes were used (black or white pixels), and in [5] binary sequences were obtained through the use of two thresholding methods: an adaptive single threshold for series from a pseudo-random number generator, and a fixed threshold, at 0.5, for series generated by logistic map difference

equations. In [8], a study of the influence of the thresholding method on LZ complexity measure is presented, considering more than one threshold (more than two symbols in resulting sequences). Unfortunately, that work does not take into account multivariate signals, whose quantization is a more complex matter.

Similarly, in [12], nine groups of signals were analysed with a pool of complexity measures, including Lempel-Ziv's one. Again, the simplest procedure to express multivariate time series of dynamical data as a symbolic sequence was used there. This procedure was the calculation of Euclidean distances between consecutive points, followed by the comparison of these distances to their median, thus yielding two-symbol sequences.

In this paper, we address multivariate signal comparison with a LZ-like method. To properly explain our approach, we first present the LZ method in Section 2. In Section 3, a new similarity measure between signals is proposed, inspired by the complexity measure defined in [1]. In Section 4, it is shown how to use this new measure as an alternative to likelihood based criteria. Also in that Section, we gather some experimental results with vector quantization of multivariate signals from a public database. These results are analysed and conclusions are presented in Section 5.

2 The Lempel-Ziv's Method

Let s_1^n be a sequence of n symbols drawn from a finite alphabet, $\mathcal{A}$, Lempel-Ziv's complexity analysis is based on the parsing of s_1^n into a minimum number of unique (with one possible exception) subsequences of symbols. Though the idea of a dictionary of subsequences is not presented in the seminal paper published in 1976 [1], we believe that it is a powerful point of view for pattern recognition. Therefore, we define a growing dictionary of subsequences s_i^j, where s_i^j stands for a substring formed by symbols from position i to position j ($i \leq j \leq n$). Thus, Lempel-Ziv's algorithm can be summarized as follows:

1. Set $k = 1$, $L = 1$ and start with a single element in the dictionary, $\{s_1\}$. Set a pointer to the first symbol of the sequence, $p_k = 1$.
2. Increase L by one, $L \leftarrow L + 1$. If $p_k + L - 1$ equals n, then take the new subsequence of length L, $s_{p_k}^{p_k+L-1}$, as the last parsed segment and stop the algorithm, otherwise...
3. Compare $s_{p_k}^{p_k+L-1}$ (a substring of length L) to every subsequence of the same length in the past[1], $s_1^{p_k+L-2}$.
4. If this search fails:
   - The subsequence is given as a new dictionary entry
   - p_k is set to the position of the next symbol, $p_k \leftarrow p_k + L$
   - k is increased by one, $k \leftarrow k + 1$
   - L is set to zero, $L \leftarrow 0$, and
   - the algorithm flow moves back to step (2).

The number of parsed substrings through this process, $C(s_1^n)$, was proposed by Lempel and Ziv as a complexity measure. Figure 1 illustrates this complexity analysis and dictionary construction for a sequence of binary symbols.

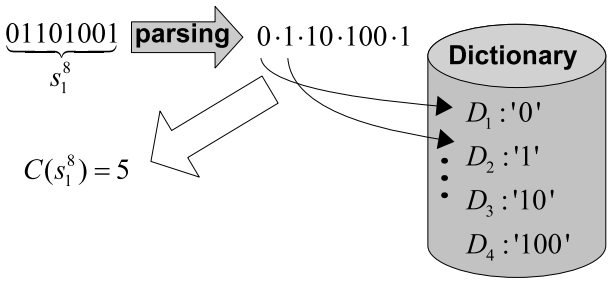

Figura 1: A very simple illustration of Lempel-Ziv's complexity analysis.

3 A new similarity measure for pattern recognition

In pattern recognition, a typical approach for signal classification can be summarized as:

- Extract features from available signals in a training database.
- Choose and adjust a probabilistic model (learning step).

[1] An alternative method was proposed in 1978 [13] to alleviate the computational burden, in which new subsequences are compared to subsequences of the same length in the dictionary.

- As for new signals, extract their features and compute their corresponding likelihood (from the learned probabilistic model). Then decide whether it is a new instance from the same source of signals or not, by comparing likelihood to a threshold.

In other words, classification is usually based on likelihood, or distances which can be regarded as a simplification of likelihood based criteria.

In this paper, we claim that compression rate may play a role similar to likelihood. Considering a lossless compression scheme, the higher the compression rate of a given signal, the greater is the number of long segments of samples that can be found more than once in the whole signal. In order to use compression rate to compare two signals, we modify the LZ algorithm to consider two series as inputs, where one of them just plays the role of "past signal", and the other is parsed.

Let x_1^m and y_1^n be two sequences of m and n symbols, both drawn from a finite alphabet, $\mathcal{A}$. Alternatively, x_1^m and y_1^n may be two sampled signals, both with samples lying in $\mathcal{R}^D$. The *Modified* Lempel-Ziv (MLZ) parsing procedure starts by searching for $y_1^{p_1}$ (initially $p_1 = 1$) inside x_1^m. If this search succeeds, p_1 is replaced with $p_1 + 1$ and a new search is done, otherwise the pointer p_1 is registered, and a new pointer $p_2 \leftarrow p_1 + 1$ is created, and the search in x_1^m for a subsequence equal/similar to $y_{p_1+1}^{p_2}$ is done; and so on. The algorithm stops when the end of y_1^n is reached. The number of segments into which y_1^n is parsed is denoted by $C_{MLZ}(y_1^n; x_1^m)$, thus indicating that x_1^m plays the role of a parameter of C_{MLZ}.
Illustration: If $x_1^6 = abbaab$ and $y_1^9 = ababbaabb$, the modified parsing process produces: $aba \cdot bbaabb$. By denoting it as a function of y, parametrized by x, we write: $C_{MLZ}(y; x) = 2$.

It is worth noting that measure $C_{MLZ}(y_1^n; x_1^m)$ depends on m, the length of "parameter" x_1^m. Parameter x_1^m can be regarded as a memory of subsequences of symbols/samples, and the longer such memory is, the more likely we are to find longer segments of y in x, even if y and x are generated by statistically independent sources. To properly compensate for this dependency of the proposed measure, C_{MLZ} on m, we make an appeal to an important statement in [1], according to which the maximum complexity measure of a sequence x_1^m, produced by an ergodic α-symbol source (α is the cardinality of $\mathcal{A}$), is $m/\log_\alpha(m)$. This upper bound gives a clue concerning the averaged length of dictionary items (sub-sequences of symbols) produced by LZ processing of x_1^m. That is to say that $m/\log_\alpha(m)$ estimates the number of segments into which x_1^m is parsed, by assuming an averaged segment length of $\log_\alpha(m)$.

As a result, we claim that the averaged length of segments of y_1^n found in x_1^m, given by $\frac{n}{C_{MLZ}(y_1^n; x_1^m)}$, is a measure of similarity between the two sequences, but it depends on m. In order to make this measure less sensitive to the length of x_1^m, we divide it by $\log(m)$ (the logarithm base is not relevant, because its change corresponds to multiplying the measure by a constant scale factor). Accordingly, we propose a new similarity criterion between sequences of symbols, given by:

$$S(y_1^n; x_1^m) = \frac{n}{C_{MLZ}(y_1^n; x_1^m)\log(m)} \tag{1}$$

and because this measure is not symmetric, i.e. $S(y_1^n; x_1^m) \neq S(x_1^m; y_1^n)$, we finally propose a symmetric criterion, given by:

$$J(y_1^n, x_1^m) = (1/2)\left(S(y_1^n; x_1^m) + S(x_1^m; y_1^n)\right); \tag{2}$$

4 Experimental Results

A straightforward approach to Lempel-Ziv analysis of sequences of numeric patterns is to vector quantize it and to perform analysis on the resulting sequence of labels (e.g. prototype labels). Though it is indeed a straightforward solution, it raises some nontrivial questions, such as the influence of number of prototypes and quantization strategy on the analysis result. In this preliminary paper, we do not discuss such difficult matters, which are postponed to future works. Instead, we empirically adjust the number of prototypes, K, through error ratio measures. Thus gathering some empirical evidences concerning the usefulness of the criterion defined in Eq. 2 for signal detection.

Experiments reported here are concerned with remote healthcare. More precisely, we analyze signals from an accelerometer attached to a subject under surveillance. These accelerometer signals are publicly available at the UCI repository [15] and was used in [16]. In our experiments, we use only signals from the belt sensor, as if it was registered by a single smartphone carried by the subject under medical surveillance. Moreover, we only gather three subsets of signals, corresponding to the following classes:

- Class 'patient falling': 5 signals
- Class 'patient lying': 5 signals
- Class 'patient walking': 5 signals

Each recorded signal corresponds to a set of 3-D vectors of regularly sampled measures in 3 orthogonal directions. More details concerning signal description and acquisition can be found in [15, 16]. Here, unlike [16], we do not use context-dependent reasoning, but a much simpler approach based on direct signal comparisons to explore LZ capabilities on signal detection. Accordingly, the only signal processing technique applied to these multivariate signals is normalization of signal power for each channel (i.e. each orthogonal direction).

The experiments are carried out as follows: a single data file, out of 15, is taken as a reference. For instance, lets assume that a signal from 'subject falling' is arbitrarily taken as a single reference for 'falling' event. Then a second signal is randomly taken, playing the role of an online recording during actual patient monitoring. These two signals are compared, providing a score.

A detection threshold is adjusted to the point where false negative and false positive rates are the same (EER), and if a score from two signals from the same class is under this threshold, a 'false negative' event is computed. On the other hand, if a score from two signals from different classes lies above it, a 'false positive' is computed.

In all independent experiments, only 4 recordings from class 'walking' were scored below threshold in both approaches, as illustrated in Figure 2, from one single but representative run of the method with quantization. Therefore, we consider this result as a stable one, because it remains unchanged for a wide range of parameter values, namely: all signal quantizations with a single K-means run, with K ranging from 8 to 23.

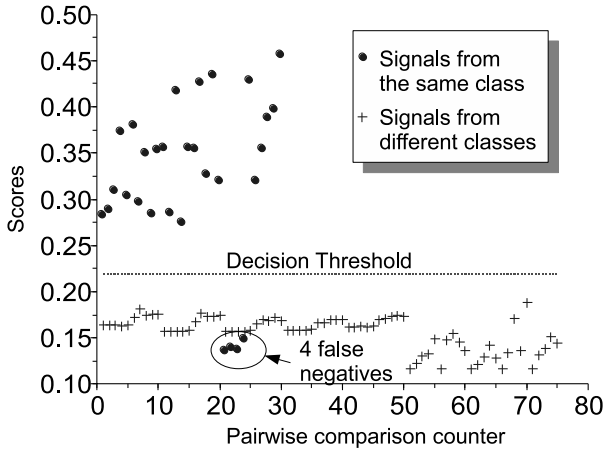

Figura 2: Comparison scores and classification of accelerometer signals — 30 cross-comparisons of signals from the same class, 75 cross-comparisons of signals from different classes, only 4 false negative results.

5 Conclusion

In this paper, the Lempel-Ziv complexity measure, usually associated to lossless compression of computer files, was used in a less common task: pattern recognition. A modified version of the LZ analysis method proposed in 1976 [1] was presented along with a new similarity measure aimed at comparing two sequences instead of computing complexity of a single one. This modified method, MLZ, uses the same simple but powerful ideas behind the original LZ method. Likewise, the new similarity measure uses the number of parsed segments instead of usual likelihood based measures used in pattern recognition.

Another contribution of this work was a method for continuously valued signal analysis through LZ algorithm, based on signal space discretization in K prototypes, through the K-Means algorithm. By seeing signals as instances of stochastic processes, detection of such signals usually relies on time signal alignment, which is a critical aspect in dynamic pattern recognition, being sometimes explicitly imposed through Dynamic Time Warping (DTW), sometimes indirectly modelled in Markov models [14]. By contrast, MLZ does not need previous time signal alignment. Therefore, in this preliminary short paper, in order to just illustrate the usefulness of the proposed tools, experimental results were presented on healthcare signal detection, which can be regarded as a typical problem to be modelled with stochastic processes. Experiments yielded promising performances, and even if comparisons to other "classical" strategies are not yet provided (under preparation), it does illustrate that the LZ-complexity inspired criterion may indeed play the role of likelihood.

For the experiments with accelerometer signals (Healthcare) consistent detection results were obtained, with errors on only 4 out of 105 cross-comparisons for a wide range of algorithm parameters. In all runs, the number of prototypes, K, was empirically tuned, for the automatic tuning of these parameters is beyond the scope of this paper. Nonetheless, we noticed that there is an optimum value for K, and studying the relationship between signal statistics and optimization could be a new interesting matter for future research.

Acknowledgments

This work was partially granted by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq). The authors also thank Fábio Prudente, from the *Instituto Federal de Educação, Ciência e Tecnologia de Sergipe* (IFES-SE) for his challenging point of view and valuable remarks on the practical applications of this work.

Referências

[1] Lempel, A., Ziv, J., 1976. On the complexity of an individual sequences. *IEEE Trans. on Inform. Theory*, IT-22, pp. 75–81.

[2] Shannon, C. E., 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27, pp. 379–423 and 623–656.

[3] Cover, T.M, Thomas, J.A. 2006. Elements of Information Theory, John Wiley & Sons, Inc.

[4] Johansen, P. 1995. Pattern recognition by data compression, in *Proceedings of the 9th Scandinavian Conference on Image Analysis*, pp. 1131–1137.

[5] Kaspar, F., Schuster, H. G. 1987. Easily calculable measure for the complexity of spatiotemporal patterns. *Physical Review A*, 36, no. 2 pp. 842–848.

[6] Batista, L. V., Meira M. M.. 2004. Texture Classification Using the Lempel-Ziv-Welch Algorithm, In *C Science Advances in Artificial Intelligence - SBIA 2004, Lecture Notes in Computer Science*, 3171, 2004, pp. 444-453.

[7] Amigó, J. M., Szczepanski, J., Wajnryb, E., Sanchez-Vives, M. V. 2004. Estimating the Entropy Rate of Spike Trains via Lempel-Ziv Compplexity. *Neural Computation*, 16, pp. 717–736.

[8] Sarlabous L., Torres A., Fiz J.A., Gea J., Galdiz J.B., Jane R. 2009. Multistate Lempel-Ziv (MLZ) index interpretation as a measure of amplitude and complexity changes, In *Conf Proc IEEE Eng Med Biol Soc.*, pp. 4375–4378.

[9] Aboy, M., Hornero, R., Abásolo, D., Alvarez, D. 2006. Interpretation of the lempel-ziv complexity measure in the context of biomedical signal analysis. *IEEE Trans Biomed Eng*, 53(11), pp. 2282–2288.

[10] Aboy, M. Cuesta-Frau, D., Austin, D., Micó-Tormos, P. 2007. Characterization of Sample Entropy in the Context of Biomedical Signal Analysis, In Proceedings of the 29th *Conf Proc IEEE Eng Med Biol Soc.*, pp. 5942–5945.

[11] Gao, Y., Kontoyiannis, I., Bienenstock, E. 2008. Estimating the Entropy of Binary Time Series: Methodology, Some Theory and a Simulation Study. *Entropy*, 10, pp. 71-99.

[12] Rapp, P.E., Watanabe, T. A. A., Faure, P., C. J. Cellucci, C. J. 2002. Nonlinear Signal Classification, *International Journal of Bifurcation and Chaos*, 12, No. 6, pp. 1273–1293.

[13] Ziv, J., Lempel, A. 1978. Compression of Individual Sequences via Variable-Rate Coding. *IEEE Trans. on Inform. Theory*, IT-24, pp. 530–536.

[14] Theodoridis, S., Koutroumbas, K. 2003. Pattern Recognition (2nd Ed.), Elsevier Academic Press.

[15] Blake, Merz, C.J. 1998. UCI repository of machine learning databases.

[16] Kaluza, B., Mirchevska, V., Dovgan, E., Lustrek, M., Gams, M. 2010. An Agent-based Approach to Care in Independent Living, In Proceedings of the First international joint conference on Ambient intelligence, pp. 177–186.